بررسی وب‌کاوی

فرشته آرمانفر
دانشکده تحصیلات تکمیلی، دانشگاه آزاد اسلامی، بروجرد
Armanfar.90@gmail.com

چکیده

وب‌کاوی[1] زمینه‌ی مهمی از داده‌کاوی[2] است که مقادیر مورد استفاده آن با استخراج دانش موردنظر از WWW بدست می‌آید. داده-کاوی در سه زمینه داده‌کاوی، متن‌کاوی و وب‌کاوی عمل کاوش را انجام می‌دهد. در این مقاله از راهکارهایی استفاده شده که مفهوم وب‌کاوی را معرفی می‌کنند. وب‌کاوی با داده نیمه‌ساخت‌یافته یا غیرساخت‌یافته سروکار دارد و خواستار استفاده خلاقانه از تکنیک-های داده‌کاوی است. فرآیند وب‌کاوی در چهار زیروظیفه‌ی منبع کلکسیون داده، پیش پردازش داده، اکتشاف الگو و تحلیل الگو گروه‌بندی شده است. وب‌کاوی را می‌توان در سه نوع متفاوت Web Content Mining، Web Structure Mining و Web Usage Mining دسته‌بندی کرد. هدف از این مقاله ارزیابی گذشته و حال وب‌کاوی است. همچنین خلاصه‌ای از روش‌ها و تکنیک-های وب‌کاوی و کاربردهای آن ارائه شده و مروری بر توسعه پژوهش‌ها و برخی از نتایج مهم پژوهش‌ها ی اخیر ارائه می‌شود.

کلمات کلیدی

Web Mining، Data Mining، Web Content Mining، Web Structure Mining، Web Usage Mining

۱- مقدمه

در عصر حاضر انسانها بیش از هر زمانی داده و اطلاعات ایجاد می‌کنند وآنرا منتشر می‌سازند. در واقع اطلاعاتی که در اختیار ما قرار دارد بسیار بیشتر از آن است که بتوان آن‌را تجزیه و تحلیل کرد. به همین دلیل، انتخاب منابع مورد نظر کاربر با این حجم زیاد اطلاعات روز به روز مشکل‌تر می شود. این مسئله دلایل مختلفی دارد، از جمله اینکه وب خیلی بزرگ و پیچیده است و رشد روز افزونی دارد. داده های وب به سرعت در حال تغییرند. با وجود اینکه وب از لحاظ اندازه به سرعت در حال رشد است و اطلاعات آن بطور مداوم تغییر می‌کند و به‌روز می-شود. وب هیچ‌گونه سازماندهی منطقی ندارد. گرچه ممکن است بخش-های کوچکی از وب سازماندهی خوبی داشته باشند اما درکل وب سازماندهی کاملا غیرساخت‌یافته دارد.

علاوه بر اینها وب کاربران متفاوتی دارد و هرکاربر ممکن است تنها بخش کوچکی از وب را مدنظر قرار دهد. در نتیجه، کاربران آنلاین مشکلات زیادی در یافتن اطلاعات موردنظر دارند. موتورهای جستجو به کاربران در یافتن منابع اطلاعاتی مرتبط کمک می‌کنند اما موتورهای جستجو دارای مشکلات زیادی هستند. یک عنوان ممکن است در صدها یا هزاران سند را وجود داشته باشد که باعث می‌شود موتور جستجو اسناد غیرمرتبط زیادی را همراه نتایج جستجو برگرداند. همچنین تعداد زیادی از اسناد که کاملا به عنوان مورد نظر وابسته هستند ممکن است کلمات‌کلیدی که حاوی آن موضوع است وجود نداشته باشد. مجموعه اطلاعات موجود در وب منابع غنی برای وب-کاوی بوجود می‌آورند که می‌تواند در جستجو و استخراج منابع اطلاعاتی به‌همراه بهبود در عملکرد موتورهای جستجو کمک کند.

در این مقاله ابتدا بررسی اجمالی برروی داده‌کاوی و وب‌کاوی می-پردازیم. ادامه مقاله به این صورت سازماندهی شده، در بخش ۳ فرآیند وب‌کاوی شرح داده شده است. در بخش ۴ طبقه‌بندی وب‌کاوی را بطور خلاصه بیان کردیم و در بخشهای۵ ،۶ و۷ مهمترین روشها وپژوهش‌ها، از گذشته تا به حال در سه ناحیه وب‌کاوی مورد بررسی قرار خواهیم داد و درپایان در بخش ۸ نتیجه گیری را ارائه داده‌ایم.

۲- بررسی اجمالی

در سال ۱۹۹۶برای اولین بار **Etzioni** واژه وب‌کاوی را بکار برد. **Etzioni** ابتدا فرضیه‌ای بیان کرد که درآن اطلاعات روی وب به اندازه کافی ساخت‌یافته بودند و همچنین طرح کلی زیروظایف وب-کاوی را بیان نمود. وب‌کاوی **Etzioni** از تکنیک‌های داده‌کاوی برای کشف و استخراج اطلاعات بصورت خودکار از اسناد **WWW** استفاده می‌کرد(۲). بطور کلی در داده‌کاوی سه نوع کاوش وجود دارد : داده-کاوی، متن‌کاوی و وب‌کاوی. چالش‌های زیادی در مسائل این سه حوزه تحقیقاتی وجود دارد (۱). دست‌یافته‌های تحقیقات داده‌کاوی از ارتباط های نادرست در حل بعضی از مسائل اصلی تحقیقات آماری مانند مسئله**old age** جلوگیری می‌کند. داده‌کاوی بطور عمده با داده ساخت‌یافته در یک پایگاه داده سروکار دارد در حالیکه متن‌کاوی بطور عمده با داده یا متن غیرساخت‌یافته سروکار دارد. وب‌کاوی در میانه قرار دارد و با داده نیمه‌ساخت‌یافته ویا غیرساخت‌یافته سروکار دارد.

آکادمی داده
dataacademy.ir

وب‌کاوی خواستار استفاده خلاقانه از تکنیک‌های داده‌کاوی و یا متن-کاوی است و رویکردهای آن مشخص است. استخراج داده‌های وب یکی از چالش‌انگیزترین وظایف داده‌کاوی وپروژه مدیریت‌داده است(۲). در اینجا تعاریفی برای داده‌کاوی و وب‌کاوی ارائه می دهیم : داده‌کاوی عبارت است از استخراج اطلاعات و دانشی که برای افراد ناشناخته است و دانش بالقوه را از تعداد زیادی داده تصادفی ناکامل ومبهم از برنامه ها استخراج می‌کند. وب‌کاوی کاربردی از تکنولوژی داده‌کاوی است که الگو های بالقوه، مفید و همچنین اطلاعات پنهان اسناد و فعالیت های وب را استخراج می‌کند. وب‌کاوی روی داده‌هایی مانند محتوای صفحات وب،اطلاعات دسترسی کاربر و هایپرلینک میان صفحات کار می‌کند برای دستیابی به ویژگی‌های ذاتی بین اشیاء داده از طریق روشهای **machine learning**، **inductive learning** و تحلیل آماری برای یافتن الگوهای بالقوه ، جالب و اطلاعات ضمنی مورد استفاده در روش‌های داده‌کاوی تمرکز می کند(۳).

۳- فرآیند وب‌کاوی

مطابق با فرضیه **Etzioni** وب‌کاوی به چهار زیروظیفه زیر تقسیم می شود:

- منبع کلکسیون داده
- پیش پردازش داده
- اکتشاف الگو
- تحلیل الگو

۳-۱- منبع کلکسیون داده

در استفاده و کاوش داده وب، بیشترین توجه منابع داده بر روی فایل-های‌ثبت[۲] وب، روی وب‌سرور است. فایلهای ثبت وب، رفتار ملاقات کنندگان وب را بصورت خیلی واضح ثبت و نگهداری می‌کنند. فایل های‌ثبت‌وب شامل **server log** ، **agent log** و **client log** است.

۳-۲- پیش‌پردازش داده

مجموعه داده بطور کلی ویژگی ناکامل بودن، افزونگی و ابهام را دارد. به این منظور برای استخراج موثر دانش، پیش‌پردازش مجموعه داده ضروری است. پیش‌پردازش می‌تواند دقت را برای داده‌کاوی فراهم نماید. پیش‌پردازش داده شامل پاکسازی داده، تشخیص کاربران، تشخیص **session** های کاربر، مسیر مکمل دستیابی و تشخیص تراکنش است. که در ادامه هریک را بطور خلاصه شرح می‌دهیم:

- وظیفه اصلی پاکسازی داده، حذف داده افزونه از فایل‌ثبت-وب است که به داده‌های مفید ارتباطی ندارد و به حوزه اشیاء داده محدود می‌شود.
- تشخیص کاربر، باید بعد از پاکسازی داده انجام شود. منظور از تشخیص کاربر، احرازهویت هرکاربر است. تشخیص کاربر می تواند با مفهوم تکنولوژی کوکی، تکنولوژی ثبت کاربر و قوانین اکتشافی کامل گردد.
- تشخیص **session** کاربر، براساس"تشخیص کاربر" انجام می شود. به همین منظور اطلاعات دسترسی‌کاربر به چند فرآیند **session** جداگانه تقسیم می شود. ساده ترین روش، استفاده از رویکرد تخمین **time-out** است. اگر بازه زمانی میان درخواست‌های صفحه از مقدار معینی تجاوز کند، مانند این است که کاربر یک **session** جدید را آغاز کرده است.
- بدلیل استفاده گسترده از تکنولوژی **cashing** صفحه و سرورهای **proxy**، مسیر دسترسی بوسیله فایل‌های‌ثبت وب‌سرور نگهداری می‌شود که ممکن است این مسیر ثبت شده مسیر دسترسی کامل کاربران نباشد. فایل‌ثبت ناکامل، روی الگوی دسترسی کاربران اثر می‌گذارد. بنابراین اضافه کردن مسیر دسترسی لازم نیست و با بکار بردن مسیر مکمل می توان از خواص توپولوژی وب‌سایت برای تحلیل صفحات استفاده کرد.
- تشخیص تراکنش، براساس تشخیص **session** کاربر است. به همین منظور تراکنش‌ها را مطابق با خواسته‌های وظایف داده‌کاوی تقسیم یا ترکیب می‌کند.

۳-۳- اکتشاف الگو

اکتشاف الگو اطلاعات موثر، جدید و دانش را با استفاده از الگوریتم کاوش استخراج می‌کند. این روش‌ها شامل تحلیل دسته‌بندی، اکتشاف قوانین وابستگی، اکتشاف الگوی توالی، تحلیل خوشه بندی و مدلسازی وابستگی است. که در ادامه هریک را بطور خلاصه شرح می‌دهیم:

- تحلیل دسته بندی ، طبقه بندی گروه ها مطابق با گروه های از پیش تعریف شده است ، که اساساً نشان دهنده پروفایل کاربر برای ایجاد گروه کاربران است.
- اکتشاف قوانین وابستگی، اساساً برای استخراج "قوانین مرتبط" از اطلاعات دسترسی در فایل ثبت پایگاه داده وب و یافتن ارتباطات پنهان داده‌ها بوسیله تحلیل پیوند برای دسترسی کاربران به صفحات وب استفاده می‌شود.
- اکتشاف الگوی توالی، از مدلی استفاده می‌کند که رابطه توالی زمانی بین مجموعه مورد تبادل را دارد. بر روی فایل-های ثبت سرور، ملاقات کردن کاربران بصورت یک سری زمانی گسسته است و این توالی ها روی رفتار کاربران تاثیرگذار است.
- تحلیل خوشه‌بندی، کاربر یا آیتم‌های داده با مشخصات مشابه را طبقه‌بندی می‌کند. برای مجموعه‌ی کاربران یا آیتم‌های داده با مشخصات مشابه یکدیگر نیز استفاده می-شود.

آکادمی داده
dataacademy.ir

- هدف از مدلسازی وابستگی، توسعه مدلی است که می تواند وابستگی های مهم بین متغییرهای مختلف در حوزه وب را شرح دهد که به تنهایی نمی تواند چارچوبی برای تحلیل رفتار کاربر فراهم کند اما توانایی بالقوه پیشگویی مصرف منابع وب را دارد.

۳-۴- تحلیل الگو

تحلیل الگو اساساً برای انتخاب الگوی مورد علاقه در مجموعه الگوهای یافته شده است که بوسیله مدل الگوریتم اکتشاف الگو انجام می‌شود. هدف یافتن مدل باارزشی مانند، قوانین و حالات مورد نظر و ساختن رابط گرافیکی کاربر با استفاده از تکنیک های بصری سازی است(۴,۳).

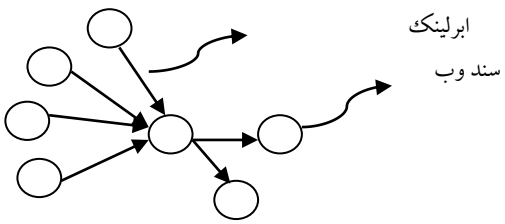
۴- طبقه بندی وب‌کاوی

وب‌کاوی براساس اینکه کدام قسمت از وب مورد بررسی است می تواند در سه ناحیه دسته بندی شود(۴).

- **Web Content Mining**
- **Web Structure Mining**
- **Web Usage Mining**

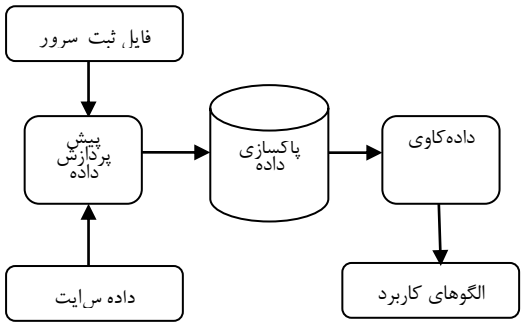
Web Content Mining فرآیند استخراج اطلاعات مفید از محتوای اسناد وب است. که شامل متن، تصاویر، صدا، ویدیو و یا رکورد های ساخت‌یافته مانند لیست‌ها و جدول‌ها می باشد. فعالیت‌های پژوهشی در این زمینه شامل استفاده از روش‌هایی از قبیل بازیابی اطلاعات **(IR)** و پردازش زبان طبیعی **(NLP)** است. قسمت اصلی کار استخراج دانش از تصاویر است - در زمینه پردازش تصویر و بینایی ماشین - استفاده از این تکنیک ها در **Web Content Mining** رشد سریعی ندارد(۵).

Web Structure Mining ساختار عمومی به‌شکل گراف دارد که شامل صفحه‌ها وب بعنوان گره‌ها و ابرلینک‌ها بعنوان لبه‌های اتصال-دهنده بین دو صفحه مرتبط است. **Web Structure Mining** می-تواند بعنوان فرآیند اکتشاف اطلاعات ساختار وب در نظر گرفته شود(۴ ,۵ ,۶).



Web Usage Mining کاربردی از تکنیک‌های داده‌کاوی برای اکتشاف الگوی کاربرد مورد نظر از داده وب است. که برای درک نیازها و پاسخ دهی بهتر به نیازهای برنامه‌های تحت وب است. نتایج **Web Usage Mining** از تعامل کاربر با یک وب‌سرور شامل فایل‌های ثبت وب، **click stream**ها و تراکنش‌های پایگاه‌داده یک یا چند سایت مرتبط، بدست می‌آید(۵). **Web Usage Mining** در سه فاز گروه-بندی می شود:

- پیش‌پردازش
- اکتشاف‌الگو
- تحلیل‌الگو



شکل (۲) : فرآیند کاوش کاربرد وب

در پیش‌پردازش، اولین رویکرد برای بازیابی داده‌ی خام از منابع وب و پردازش داده است. که بطور خودکار داده خام را تغییر شکل می‌دهد. اکتشاف الگو، بر طبق پیش‌پردازش داده، دانش را کشف می‌کند و از تکنیک های کشف دانش مانند **machine learning** و روال‌های داده‌کاوی استفاده می‌کند. تحلیل الگو، فرآیند بعدی از اکتشاف الگو است که صحت الگو را روی وب بررسی می‌کند. همچنین چگونگی پیاده‌سازی الگو در وب را برای استخراج اطلاعات از روی نتایج جستجو یا استخراج دانش بررسی می‌کند (۴).

۵- Web Content Mining

مقادیر **Web Content Mining** با اکتشاف اطلاعات یا دانش مفید از روی محتوای صفحه وب بدست می‌آید. **Margaret H. Dunham** بیان کرد **Web Content Mining** از روی اعمال انجام شده بوسیله موتورهای جستجوی پایه می‌تواند گسترش یابد. توسعه اخیر داده‌کاوی چند رسانه‌ای محدوده‌ی وسیعی از منابع وب را دربر می‌گیرد که شامل تصویر، صدا، ویدیو و غیره است(۶). از آنجا که تفسیر تصاویر روی وب بطور عادی ممکن نیست و همچنین تصاویر وب بدلیل زیاد بودن مفاهیم در عکس و منحصربه‌فرد بودن ادراک بشر، با استفاده از توصیف‌گر معنایی بخوبی تفسیر نمی‌شوند. درک کاربران مختلف از یک تصویر بندرت مشابه هم است. این مسائل به کاربردهای ابزار بازیابی تصویر مبتنی بر کلمات کلیدی و همچنین روش‌های سنتی مبتنی بر متن که عمل سنگین بارگذاری تصاویر را نمی‌توانند انجام دهند، محدود می‌شود. بنابراین مفهوم بازیابی تصویر مبتنی بر محتوا [۴]

آکادمی داده
dataacademy.ir

(CBIR) بوجود آمد. بازیابی تصویر مبتنی بر محتوا (CBIR) تلاش می کند تا فرآیند شاخص‌گذاری یا تفسیر تصاویر را در پایگاه داده تصویر بطور خودکار انجام دهد. بازیابی تصویر مبتنی بر محتوا (CBIR) از محتوای بصری تصاویر مانند رنگ، شکل و بافت برای نمایش و شاخص‌گذاری تصویر استفاده می نماید. این مفاهیم با استفاده از بردارهای ویژگی چندبعدی استخراج شده و در پایگاه‌داده شاخص‌گذاری شده‌اند. وقتی تصویری بعنوان پرس‌وجو ورودی برای بازیابی اطلاعات داده می‌شود. بردارهای ویژگی استخراج می‌شوند و سپس تصاویر با بردارهای ویژگی مشابه، از پایگاه داده بوسیله مقایسه بازیابی می‌شوند. طرح شاخص‌گذاری روش موثری برای بازیابی تصاویر از پایگاه‌داده را فراهم می‌کند. در واقع علامت‌گذاری تصاویر خسته کننده و پرهزینه است و راه حل مناسبی نیست. برای بررسی این مسئله، **H.J.Zahang** طرح انتشار کلمات کلیدی پیشرفته آماری را پیشنهاد داد. که چارچوبی برای تفسیر خودکار تصاویر درون پایگاه-داده‌ها است. بطور معمول موتورهای جستجو دقت کمی در پاسخ به پرس‌وجو دارند و ممکن است بسیاری از صفحات وب غیرمفید را بازیابی کنند و بعضی از صفحات خیلی مهم را نادیده بگیرند. **Ricardo Campos** وهمکاران مشکلات سلسله مراتب خوشه‌بندی وب را مطالعه کردند و معماری موتور **meta-search** را که **WISE** نامیده شد پیشنهاد دادند که بطور خودکار خوشه هایی از صفحات وب مرتبط می ساخت و مفهوم پرس‌وجو را دربر داشت. این خوشه‌ها سپس بطور سلسله مراتبی سازمان یافتند. مهدی حسینی و حسن ابوالحسنی یک **Query-URL co-clustering** جدید برای یک وب-سایت پیشنهاد دادند که برای ارزیابی معماری اطلاعات و ساختار لینک مفید است. در مرحله اول تمام پرس‌وجوها و **URL** های مرتبط در وب سایت خاص از فایل‌ثبت پرس‌وجو گردآوری می‌شوند و بعنوان یک گراف دو بخشی،که یک طرف آن برای جستجوها و یک طرف آن برای **URL**ها است ارائه می‌شود. در مرحله‌ی بعد یک خوشه‌بندی بدون محتوا برای خوشه‌ی پرس‌وجوها و خوشه **URL**ها بطور همزمان بکار گرفته می‌شود. سپس در مرحله آخر براساس اندازه‌گیری اطلاعات، خوشه‌های **URL** ها و پرس‌وجوها به‌ترتیب برای ارزیابی ساختار لینک و معماری اطلاعات بکار می‌روند. بازیابی محتوای وب از وظایف معمول است و نیاز به توسعه ابزار خودکار و مورد نظر کاربر، برای ارائه اطلاعات مشخص و بهبود نتایج **Web Content Mining** با استفاده از تشخیص و حذف لینک‌های افزونه دارد. این موضوع امروزه چالش اصلی تحقیقات وب‌کاوی است(۲,۷).

۶- Web Structure Mining

مقادیر **Web Structure Mining** از کشف و مدلسازی ساختار لینک وب بدست می‌آید. ابزار بازیابی اطلاعات وب تنها از متن موجود در صفحات وب استفاده می‌کند و از اطلاعات مفید درون لینک‌های وب چشم‌پوشی می‌کند. اهداف **Web Structure Mining** ایجاد ساختار خلاصه برای وب‌سایت‌ها و صفحات وب است. تمرکزاصلی**Web Structure Mining** روی اطلاعات لینک است. الگوریتم اصلی برای تحلیل ابر لینک **HITS** نام دارد. **Kelenberg** مفهوم **hub**ها (صفحاتی که به صفحات زیادی ارجاع می کنند) و **authoritie**ها (صفحاتی که مورد ارجاع صفحات زیادی قرار گرفته اند) را معرفی کرد. او مجموعه ابزار الگوریتمی برای استخراج اطلاعات از ساختارهای لینک ایجاد کرد. مسئله اصلی ایجاد چارچوبی برای پالایش مباحث مورد جستجو از طریق اکتشاف اطلاعات معتبر منابع است. **Furnkaranz** بیان کرد که وب ممکن است بعنوان گراف جهت‌دار بنظر برسد که گره‌های آن اسناد وب و لبه‌های آن ابر لینک-های بین اسناد است(۶). استفاده از ساختار گراف برای بهبود عملکرد بازیابی و دقت در دسته‌بندی است. بسیاری از موتورهای جستجو از ویژگی‌های گراف در رتبه بندی نتایج پرس‌وجو استفاده می نمایند. رشد مستمر در اندازه و استفاده از اینترنت مشکلاتی در جستجوی اطلاعات بوجود آورده است. **Fang** و **Sheng** صفحه پورتال یک وب سایت را طراحی کردند. آن‌ها سعی کردند حداکثر بهره‌وری، اثربخشی و کاربرد از صفحه پورتال وب‌سایت با انتخاب تعداد محدودی ابر لینک از مجموعه بزرگ داشته باشند و براساس روابط میان ابر لینک‌ها یک رویکرد مکاشفه‌ای برای انتخاب ابرلینک پیشنهاد دادند که **link selector** نام گرفت. که بجای خوشه بندی الگوی ناوبری کاربران، آنرا با فاصله اقلیدسی اندازه گیری می نمایند. **Nacim Fateh Chikhi**و همکاران نشان دادند که هم‌ارزی بین **HITS** و تحلیل اجزای اصلی**T** تکنیک شناخته شده‌ای برای کاهش ابعاد است. آنها با استفاده از تکنیک‌های کاهش ابعاد[۵] مختلف **(DRTS)** ساختارهای ضمنی پنهان را در ابرلینک‌های وب بررسی می‌کنند. همچنین چهار تکنیک کاهش ابعاد **(DRTS)** زیر را که برای وظایف **Web Structure Mining** است مقایسه می کنند. تحلیل اجزای اصلی **(PCA)** ، ماتریس فاکتورگیری غیر منفی **(NMF)**، تحلیل اجزای مستقل **(ICA)** و پروژه تصادفی **(RP)**. آنها روی مجموعه داده‌ها آزمایش انجام دادند و نشان دادند که ماتریس فاکتورگیری غیرمنفی **(NMF)** رویکرد امید بخشی برای تحلیل ساختار وب است و نسبت به روشهای دیگر برتری دارد. علاوه بر رتبه بندی جستجو، ابر لینک ها همچنین برای یافتن اجتماعات وب مفید هستند. یک اجتماع وب مجموعه ای از صفحات وب است که روی یک موضوع یا مطلب خاص تمرکز دارد(۲,۷).

۷- Web Usage Mining

مقادیر **Web Usage Mining** از طریق درک رفتارکاربر در تعامل با وب یا وب‌سایت بدست می‌آید. یکی از اهداف آن بدست آوردن اطلاعاتی است که به سازماندهی وب‌سایت یا تطبیق بهتر درخواست کاربر کمک کند. مدل **Web Usage Mining** نوعی از کاوش‌های فایل ثبت سرور است. هدف آن گرفتن اطلاعات دسترسی کاربران از فایل‌های ثبت برای ساخت سایت‌هایی است که می‌توانند کاملاً خودشان را با درخواست‌های کاربران تطبیق دهند، که خدمات بهتر به

آکادمی داده
dataacademy.ir

کاربران و مزایای اقتصادی بیشتر را در پی دارند. تعداد زیادی ابزار تحلیل فایل‌ثبت وب برای استخراج سوابق ذخیره شده در فایل‌ثبت صفحات وب در دسترس قرار دارد. محتویات ذخیره شده فایل ثبت اطلاعات مفید زیادی از قبیل **URL**، آدرس **IP**، زمان و از این قبیل اطلاعات را داراست. اکتشاف و تحلیل اطلاعات فایل‌ثبت کمک به یافتن مصرف کنندگان بالقوه‌تر، صفحات وب محبوب تر و غیره می‌کند. که می تواند در سازماندهی وب سایت برای دسترسی سریعتر و راحت‌تر مصرف کنندگان، ارتقاء و راهبری، جذب تبلیغات بیشتر بوسیله آگهی‌های هوشمند و نظارت موثر بر وب‌سایت موثر باشد. بیشتر داده‌های مورد استفاده برای کاوش وب‌سرورها، **proxy** سرورها یا سرورهای پایگاه داده جمع آوری می‌شوند که تمام آنها داده‌های **noisy** هستند. چون وب‌کاوی به **noise** حساس است بنابراین روشهای پاکسازی داده در اینجا ضروری است. **J.Srivastava** و **R.Cooley** پیش پردازش داده را در چند زیر وظیفه دسته بندی نمودند و سپس آن‌ها را بررسی کردند و به این نتیجه دست یافتند که حاصل پیش‌پردازش باید داده‌ای باشد که اجازه‌ی تشخیص الگوی گذر کاربر را بصورت **page view**ها، **session**ها و **click stream**ها - **click stream**ها بدهد. مورد توجه خاص هستند زیرا آنها اجازه تشخیص الگوهای ناوبری کاربران را می‌دهند. مدل‌های مارکوف کاربرد گسترده‌ای برای مدل کردن رفتارهای ناوبری کاربران در وب‌سایت‌ها دارند. کیفیت **Web Usage Mining** در اکتشاف دانش علاوه بر داده به الگوریتم نیز بستگی دارد. **Yu-Hui Tao** و همکارانش در تحقیقاتی که انجام دادند یک منبع داده جدید کشف کردند و آنرا بررسی تعمدی داده[۶] **(IBD)** نامیدندکه برای بهبود اثر بخشی کاربردهای **Web Usage Mining** است. **IBD** یک طبقه‌بندی از عملیات بررسی آنلاین، از قبیل **copy**، **scroll**، یا **save as** است که در فایل‌های ثبت نگهداری نمی‌شوند. در نتیجه، اهداف پژوهش‌ها برای ایجاد درک پایه **IBD** منجر به قبول آسان پژوهش‌هاو آزمایش ها در **Web Usage Mining** می‌شود. اخیراً تعدادی الگوریتم **Web Usage Mining** برای کاوش رفتار ناوبری کاربر پیشنهاد شده است. "روش تفکیک" یکی از ابتدایی ترین روش‌های خوشه بندی مورد استفاده در **Web Usage Mining** است. سیستم‌های توصیه‌گر مبتنی بر وب، در هدایت کاربر به صفحات مقصد در وب‌سایت ها خیلی مفید واقع می‌شوند. سیستم‌های توصیه‌گر **Web Usage Mining** برای پیشگویی قصد رفتارهای ناوبری کاربران پیشنهاد شده‌اند. پیش بینی قصد وتغییر حالت کاربر در آینده براساس داده **click stream** کاربر است. **M.Jalali** و **N.Mustapha** و همکاران یک مدل برای پیشگویی آنلاین از طریق سیستم **Web Usage Mining** توسعه دادند (۸) و رویکردی برای طبقه بندی الگوهای ناوبری کاربر برای پیشگویی قصد کاربر در آینده پیشنهاد دادند(۲,۷).

۸- نتیجه گیری

در این مقاله یک جمع‌آوری درباره پژوهش‌های انجام شده در زمینه وب‌کاوی درگذشته و حال ارائه کرده‌ایم. که برپایه سه نوع وب‌کاوی در **WWW** است. همچنین فرآیند وب‌کاوی را در چهار زیروظیفه شرح داده شده است. داده‌ی وب بطورعمده نیمه‌ساخت‌یافته یا غیرساخت-یافته است. که به دلیل ناهمگونی و بدون ساختار بودن داده وب، کشف خودکار اطلاعات دارای چالش‌های زیادی در تحقیقات است. امروزه وب‌کاوی ناحیه پرثمری برای پژوهش است.

مراجع

[1] Zhang, Q., Segall, S., "*Web mining: a survey of current research,Techniques, and software*", in the International Journal of Information Technology & Decision Making Vol.7, no.4, 2008

[2] Singh, B., Singh, H., "*Web Data Mining Research:ASurvey*", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2010

[3] Mei, Li, Cheng, Feng ,*"Overview of WEB Mining Technology and Application in E-commerce*", IEEE International Conference on Computer Engineering and Technology (ICCET), Vol.7, 2010, pp. 277-280

[4] Sharma, K., Shrivastava, G., Kumar, V., *" Web Mining: Today and Tomorrow"*, IEEE International Conference on Electronics Computer Technology (ICECT), Vol 1, 2011, pp. 399-403

[5] Ramakrishna, M. , Gowdar, L., Havanur, M., *" Web Mining: Key Accomplishments,Applications and Future Directions"*, IEEE International Conference on Data Storage and Data Engineering (DSDE), 2010, pp.187-191,doi:10.1109/DSDE.2010.53

[6] Furnkranz, J., *"Web structure mining Exploiting the graph structure of the worldwide web"*, OGAI-J. 21, 2002, pp. 17–26

[7] Kosala, R., Blockeel, H., *"Web mining research: A survey"*, ACM, SIGKDD:SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, Vol. 2, 2000

[8] Jalali, M., Mustapha, N., *" WebPUM: A Web based recommendation system to predict user future movements"*, in international journal Expert Systems with Applications 37, 2010, pp. 6201-6212

زیرنویس‌ها

۱ Web Mining
۲ Data Mining
۳ Log Files
٤ Content Based Image Retrieval
٥ Dimensionality Reduction Techniques
٦ Intentional Browsing Data

آکادمی داده
dataacademy.ir